Journal of Computer Applications
计算机应用,2013,33(9):2493 -2496
ISSN 1001-9081
CODEN JYIIDU
2013-09-01
http://www.joca.cn

文章编号:1001-9081(2013)09-2493-04
doi:10.11772/j.issn.1001-9081.2013.09.2493

Web 信息整合中的数据去重方法

刘雪琼,武　刚*,邓厚平
(北京林业大学 信息学院,北京 100083)
(*通信作者电子邮箱 wugang@bjfu.edu.cn)

摘　要:针对现有数据去重方法中存在的时间效率和检测精度低的问题,结合 Web 信息整合的特点,提出一种逐级聚类的数据去重方法(SCDE)。首先通过关键属性分割和 Canopy 聚类将数据划分成小记录集,然后精确检测相似重复记录,并提出基于动态权重的模糊实体匹配策略,采用动态权重赋值,降低属性缺失对记录相似度计算带来的影响,并对名称的特殊性进行处理,提高匹配准确率。实验结果显示:该方法在时间效率和检测精度上均优于传统算法,其中准确率提高 12.6%。该方法已应用于林业黄页系统中,取得了较好的应用效果。

关键词:Web 信息整合;相似重复记录;动态权重;模糊实体匹配
中图分类号:TP311.13　　**文献标志码**:A

Data deduplication in Web information integration

LIU Xueqiong, WU Gang*, DENG Houping
(*College of Information, Beijing Forestry University, Beijing* 100083, *China*)

Abstract: Since traditional data dedupliation methods are of low time efficiency and detection accuracy, a Stepwise Clustering Data Elimination (SCDE) method was presented based on the features of Web information integration. Firstly the whole record set was divided into sub-sets using both key attributes division and the Canopy clustering technique, and then the similar records in each sub-set were accurately eliminated. A fuzzy entity matching strategy based on dynamic weight was proposed to accurately eliminate the duplicate records, which reduced the influence of missing attribute on record similarity calculation, and the name of company was especially treated to improve the matching accuracy. The results show that the method is superior to traditional algorithms in time efficiency and detection accuracy, and the precision is improved by 12.6%. The method is applied in forestry yellow page system and performs well.

Key words: Web information integration; approximately duplicate record; dynamic weight; fuzzy entity matching

0　引言

在对海量、异构、多源的 Web 信息进行整合过程中,存在大量相似重复记录[1-2]。由于“Garbage in, garbage out”,需要对这些记录进行清洗,即数据去重。数据去重过程中需要解决两个关键问题[3]:一是缩小搜索空间,二是相似记录的匹配。解决第一个问题的传统方法大都基于排序-合并的基本思想,例如滑动窗口[4]和优先队列[5]等,但由于字符位置敏感性不能保证相似的记录排在邻近的位置,导致其不能取得很好的效果。一些研究人员针对上述问题,提出将聚类技术用于重复记录检测。例如文献[6]改进了基于密度的聚类(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)算法,文献[7]将记录映射成 Q-gram 空间中的点后采用层次聚类实现检测。聚类方法在准确率和召回率等衡量指标上均有一定提高,但在 Web 信息整合中,数据量十分庞大,时间效率仍是实际应用中的瓶颈问题。

针对第二个问题,常用的相似记录匹配算法有基本字符串匹配算法、编辑距离算法、Q-gram 算法、Smith-Waterman 算法以及基于它们的一些改进算法。这些算法较为成熟,在不同领域证明了其适用性,但在 Web 信息整合环境下的准确率并不高。Web 信息整合中的数据去重相对于一般数据去重而言有其特点,由于记录通常来自不同的数据源,而不同数据源对记录存储的目的有不同的侧重,这样致使某些记录的某些属性可能会缺失,进而导致相似记录的匹配出现较大偏差。针对相似重复记录清洗中的两个关键问题结合 Web 信息整合的特点,本文提出一种逐级聚类的数据去重(Stepwise Clustering Data Elimination, SCDE)算法,并在精确去重阶段,提出基于动态权重的模糊实体匹配策略。

1　逐级聚类的数据去重方法

本文提出的逐级聚类数据去重算法,首先由专家利用领域知识人工选定关键属性,依据其对记录集进行互无交叉的分割;然后借用 Canopy 聚类思想,使用一种开销小的算法对记录粗聚类;最后并行地在子记录集中通过两两比较精确去重相似记录,大大缩小搜索空间,降低计算量,提高了时间效率。此外,在子记录集中精确检测相似记录时,提出基于动态权重的模糊实体匹配策略,采用动态权重赋值的方法,降低了属性空值对相似记录匹配造成的影响,并对名称的特殊性进行处理,提高匹配准确率。

1.1　关键属性分割

由于 Web 信息资源丰富,一条记录往往由若干属性值组成,属性描述了实体记录的特征,但在众多属性描述记录特征时的重要性是不同的,至少有一个关键属性对相似记录匹配起决定性作用(当该属性值相同时,实体记录才有可能相

收稿日期:2013-03-19;**修回日期**:2013-05-05。　**基金项目**:中央高校基本科研业务费专项基金资助项目(BLYX200928)。
作者简介:刘雪琼(1986 -),女,河北石家庄人,硕士研究生,主要研究方向:Web 信息整合、数据挖掘;　武刚(1962 -),男,北京人,教授,博士生导师,博士,主要研究方向:电子商务、信息整合、数据挖掘;　邓厚平(1989 -),男,湖北荆门人,硕士研究生,主要研究方向:信息集成。

同)。关键属性划分采用了一种分组的思想[6],由用户结合特定应用领域的知识[8]指定关键属性进行数据分割。具体选择过程中应兼顾如下三个方面:

1)关键属性值应为离散型、可枚举的;

2)关键属性值的枚举数量应尽量大,这样聚类形成的子记录集越小,有利于后续重复记录检测;

3)关键属性值的平均字符数量应尽量小,这样聚类的计算量越小。

用户人工选定 n 个具有区分度的关键属性或属性的特定部分集 $KeyList\{key_1,key_2,key_3,\cdots\}$,并按照重要程度排序;然后依据给出的关键属性,将完整记录集划分成若干个互不相交的子记录集。其具体策略是:

1)首先选取关键属性 $Key1$,枚举出它的所有 N 个取值,取值相同的记录聚为一类,则可将大记录集划分成 N 个不相交的小记录集。例如使用省份分割黄页数据,我国共有 34 个省级行政区划,这样就可以把记录分割成 34 个互不相交的集合。

2)若划分后的记录集仍十分庞大,则选择 $Key2$,对这些记录集再次分割。

3)若记录集仍不满足数据量要求,可重复第 2)步,直到记录集划分比较合理为止。

算法 1 关键属性分割。

输入 总记录集 S,阈值 a,关键属性集 $KeySet\{key_1,key_2,key_3,\cdots\}$。

输出 互不重叠的子记录集 $S_1,S_2,S_3,\cdots,S_n$。

```
SubSet(S, a, key1)
For each xi in S                          //从数据集 S 中依次取记录 xi
    根据 xi 在 key1 上的值,找到相应的子记录集 Si
    Si = Si∪{xi}
For each Si in(S1, S2, S3, ..., Sn)
    If(|Si| > a) then
        SubSet(Si, a, key2)               //继续分割
Return(子记录集 S1, S2, S3, ..., Sn)
```

1.2 Canopy 聚类算法

在此阶段借用了文献[9]提出的一种用于处理大型数据库的新型聚类技术,即 Canopy 聚类,对上一步记录集进行粗聚类。其主要思想是首先以某一条记录为中心,用一种开销较小的相似度计算方法,本文采用了信息系统中使用较多的倒排检索方法[10],利用共有词比例作为记录间距离,把数据高效地划分成可重叠的子集(即 Canopy),然后在 Canopy 子集中精确清洗重复记录。这样,只需要对 Canopy 子集中内的记录进行精确的去重,从而大大减少了传统聚类算法中对所有数据进行比较的计算量,此外,Canopy 聚类允许产生重叠的子集起到了消除孤立点作用,增加了算法的容错性。

图 1 为 Canopy 的示意图[11],在分区后的子记录集中随机选择一个数据记录 A,所有与 A 间距离小于某一阈值 $T1$(实线内)的记录标记为同一 Canopy 子集,把与 A 间距离小于某一阈值 T_2(虚线内)的记录从记录集中删除,其中 $T_1\geqslant T_2$。Canopy 中心 B、C、D、E 的形成过程同 A。

算法 2 Canopy 聚类。

输入 记录集 S,距离阈值 T_1 和 T_2($T_1\geqslant T_2$)。

输出 聚类集合 CanopySet($C_1,C_2,C_3,\cdots,C_n$)

```
while(存在未标记的数据点){
    随机选择一个没有标记的记录对象 R;
    把 R 看作一个新的 Canopy Ci 的中心;
    遍历计算其他记录对象 Ri 与 R 间距离 DRRi;
    若 DRRi < T1,则把该记录对象标记为 Ci 子集记录;
    若 DRRi < T2,则把该记录从总记录集中删除;}
```

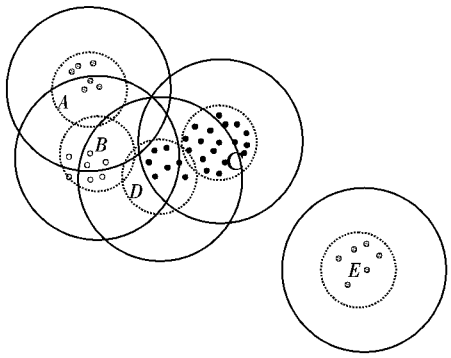

图 1 Canopy 示意图

1.3 精确去重阶段

在每个 Canopy 中通过两两比较记录间的相似度来精确去重记录集。由于记录来自不同 Web 信息源,信息侧重不同,属性可能不完整,传统的相似度匹配方法准确率较低。为了减小属性缺失对实体匹配带来的影响,本阶段提出一种基于动态权重的模糊实体匹配策略,在第 2 章将作详细阐述。对于检测出的相似记录,只能保存相似记录中的一条并删除其他记录。提供三种方法来处理检测出的相似重复记录:1)择其一保留,清除其他重复记录;2)合并,即把每条重复记录看作完整记录的一部分,将其合并作为一条包含更完整信息的新记录;3)由专家根据匹配结果判断取舍。可依据具体情况的要求选择不同策略。

2 基于动态权重的模糊实体匹配策略

针对 Web 信息整合的特点,提出一种基于动态权重的模糊实体匹配策略,针对事物名称缩写、简称等问题的特殊性导致的匹配问题,本文提出模糊名称匹配策略,提高了匹配精度。

2.1 基于动态权重的记录相似度计算

如前文所述,记录间相似度的计算方法有多种,其中编辑距离算法由于计算简单、效果明显,最为常用。所以,本文基于编辑距离来阐述动态权重的记录相似度计算方法。设记录集 $C=\{r_1,r_2,\cdots,r_n\}$,r_i 表示记录集的第 i 条记录,属性集 $P=\{p_1,p_2,\cdots,p_m\}$,p_k 表示记录的第 k 个属性,r_{ik} 表示记录 r_i 第 k 个属性的值。对于任意两条记录 r_i 和 r_j 在第 k 个属性上的值分别为 r_{ik} 和 r_{jk},令 $D(r_{ik},r_{jk})$ 表示其编辑距离,则其相似度 $Sim(r_{ik},r_{jk})$ 可以用式(1)表示:

$$Sim(r_{ik},r_{jk})=\begin{cases}1-\dfrac{D(r_{ik},r_{jk})}{\max(|r_{ik}|,|r_{jk}|)} & D(r_{ik},r_{jk})\neq\infty\\ 0 & D(r_{ik},r_{jk})=\infty\end{cases}\tag{1}$$

记录间相似度定义为记录所有对应属性相似度的加权平均。由于记录每个属性对于检测重复记录的贡献不同,所以不同属性按贡献大小赋予不同权值是一种合理且成熟的方法。假设各属性贡献比例为 $c_1:c_2:\cdots:c_n$,全部属性权值之和为 1。用 W_k 表示属性 P_k 的权重,计算公式:$w_k=c_k/\sum c_i$。任意两条记录间相似度 $Sim(r_i,r_j)$ 可以用式(2)表示

$$Sim(r_i,r_j)=\sum_{k=1}^{n}Sim(r_{ik},r_{jk})*w_k\tag{2}$$

在进行相似度计算过程时,若记录的某条属性信息缺失,则该属性相似度为 $Sim(r_{ik},r_{jk})=0$,这使得记录相似度的总值下降,$Sim(r_i,r_j)$小于阈值,两条记录被视为不同。然而,r_1 与 r_2 可能是相似重复记录。这样的固定权重计算方法由于属性缺失,给重复记录识别带来了很大影响。本文提出一种动态权值的计算方法。

设用 Val_k 表示属性 k 的有效性,定义

$$Val_k=\begin{cases}1, & \text{属性 } k \text{ 有效}\\ 0, & \text{属性 } k \text{ 无效}\end{cases}$$

如果某一属性缺失或无效,则全部属性的权重随之动态变化,减少因属性缺失而带来的影响。将属性有效性考虑进来,W_k 可表示为 $w_k=(Val_k*c_k)/\sum(Val_i*c_i)$。由于动态权重始终保证了记录对中全部有效属性的权值和为 1,从而降低了属性不完整记录对相似度计算的影响。综上,基于动态权重的记录间相似度可以用式(3)表示:

$$Sim(r_i,r_j)=\sum_{k=1}^{n}Sim(r_{ik},r_{jk})*\frac{Val_k*w_k}{\sum_{i=1}^{n}Val_i*w_i} \tag{3}$$

2.2　模糊实体名称匹配

在很多领域的记录中,最重要的属性之一是事物的名称,黄页中的企业名称即为典型的一例。然而,各种缩写和简称使得这一重要属性的匹配准确率并不高,常常难以识别不同的表达是同一企业。这种情况下,有必要对企业名称进行一些特殊的处理。

按一般惯例,企业名称由行政区划 + 字号 + 行业 + 组织形式四部分依次组成。例如:“湖北省金林木业有限公司”。各部分具有不同的权重和特点,需要分别计算,给出了企业名称模糊匹配策略,如图 2 所示。

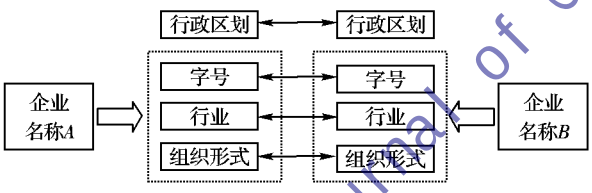

图 2　企业名称模糊匹配

首先,利用全国行政区划数据库和分词软件对企业名称的四个组成部分进行分割。通过查询全国行政区划数据库可以唯一确定“行政区划”,利用分词软件对企业名称剩余部分进行分词。由于分词软件的基础词库不可能包含较全面的专业领域词汇,因此需要进行词库添加,本文使用文献[12]的整理得到的林业专业词库,有效识别林业企业的“行业”部分。将“行政区划”和“行业”有效分割后,企业名称的四个组成部分也就分割开了。

然后,对四部分分别进行相似度计算。行政区划指本企业所在的县级以上行政区划的名称或地名。相同行政区划中的两条记录不一定为同一实体,但不同行政区划的两条记录一定不为同一实体,当 A 和 B 位于不同行政区划时,可以视为其距离无穷大,则其行政区划相似度 $Sim_{reg}(A,B)=0$,此时 A 和 B 不可能为相似重复记录,即 $Sim(A,B)=0$;当 A 和 B 位于同一行政区划时,根据其粒度不同,可预设不同的相似度 Sim_{reg},再同其他三个组成部分加权计算,则企业名称 A、B 的相似度计算方法如式(4)表示:

$$\begin{aligned}Sim(A,B)=&\alpha Sim_{reg}(A,B)+\beta Sim_{name}(A,B)+\\&\gamma Sim_{bus}(A,B)+\delta Sim_{org}(A,B)\end{aligned} \tag{4}$$

其中:α、β、γ 和 δ 分别为企业名称四部分的权重,设其贡献比值为 1∶1∶1∶1,$0\leqslant\alpha,\beta,\gamma,\delta\leqslant1$,$\alpha+\beta+\gamma+\delta=1$;$Sim_{name}(A,B)$、$Sim_{bus}(A,B)$、$Sim_{org}(A,B)$分别为字号、行业、组织形式的相似度。$Sim_{name}(A,B)$、$Sim_{bus}(A,B)$、$Sim_{org}(A,B)$三者计算方法相同,以字号相似度为例,基于式(1)可表示为:

$$Sim_{name}(A,B)=\begin{cases}1-\dfrac{D(A_{name},B_{name})}{\max(|A_{name}|,|B_{name}|)} & D(A_{name},B_{name})\neq\infty\\ 0 & D(A_{name},B_{name})=\infty\end{cases} \tag{5}$$

最后,可推得企业名称 A、B 的相似度计算公式表示为:

$$Sim(A,B)=\begin{cases}\alpha Sim_{reg}(A,B)+\beta Sim_{name}(A,B)+\gamma Sim_{bus}(A,B)+\\ \quad\delta Sim_{org}(A,B) & Sim_{reg}(A,B)\neq0\\ 0 & Sim_{reg}(A,B)=0\end{cases} \tag{6}$$

3　算法复杂度分析

逐级聚类数据去重算法的第一层关键属性分割实现同查找算法相似,时间复杂度为 $O(n)$。第二层 Canopy 粗聚类采用倒排检索,假设记录集有 n 条记录,每条记录有 v 个词,使用 V 个词检索全体记录,则算法的时间复杂度可表示为 $O(nv^2/V)$,由于 v 远远小于 V,本层聚类时间复杂度会小于 $O(n)$。因此逐级聚类算法的计算量主要在于第三层精确去重。Canopy 聚类后,假设创建了 c 个 Canopy 子集,平均每条记录被标记进 r 个 Canopy 中,所以得到每个 Canopy 中有 rn/c 个记录。因为只需两两比较 Canopy 内记录间相似度,所以最多需要计算 $O(c(rn/c)^2)=O(r^2n^2/c)$次。但由于 r 是一个极小的量,因此 r^2/c 远小于 1。这时本文算法的复杂度已明显优于排序-合并算法。此为串行情况下的时间复杂度,由于在第三层采用并行处理,理论上时间复杂度应为上述值的 $1/c$,即为 $O(r^2n^2/c^2)$,但并行处理实际运行时间和效率受处理器核数、程序运行时内存占用情况等多种因素影响,常常不能达到理论值。

4　实验验证

4.1　实验设置

算法实现使用 C#编程语言,Visual Studio 2010 作为开发工具,微机环境为 2.3 GHz 双核 CPU,2 GB 内存。

实验整合了来自 36 个林产品网站获得的 60 万条林业企业数据,记录属性包括林业企业名称、企业地址、联系电话、所属省市、邮政编码、电子邮件、电子地址、主营业务、详细信息等 9 个属性。设定各属性贡献比值为 9∶3∶2∶2∶1∶1∶1∶1∶1。随机抽取其中 800 条记录作为测试集,人工识别得到的重复数据大约占 24.8%,包含两条以上的重复记录为 198 条,其中最多的重复记录包含 12 条记录。实验步骤如下:

1)关键属性分割。人工选择关键属性集为{省,市},并对全体记录分割。

2)Canopy 聚类。采用倒排检索方法,共有词的比例作为

距离阈值。

3)精确去重。利用. net 4 中的任务并行库(Task Parallel Library, TPL)对 Canopy 中的记录进行数据并行处理,采用中国科学院的 ICTCLAS 进行分词,采用基于动态权重的模糊实体匹配策略计算记录间相似度,将识别出的重复记录合并作为一条新的记录,删除其余记录。

4.2 实验结果

实验选择 DGHS 算法[7]作为对比算法,分别从准确率、召回率和运行时间等方面进行了对比实验,成功完成了林业 Web 黄页信息整合与数据去重任务。

1)准确率和召回率对比。

实验采用测试集的 800 条数据进行测试。逐级聚类过程中有三个参数待确定:Canopy 聚类中距离阈值 T_1、T_2 和记录间相似度 T_{sim}。其中,T_1 和 T_2 通过精确聚类时的匹配对的计算量来确定; T_{sim} 的确定依据 $F1$ 值确定。从表 1 和图 3 可以得到,当 $T_1=0.3$, $T_2=0.25$ 的时候,精确聚类的计算量达到最小,而当 $T_{sim}=0.74$ 时, $F1$ 值最高,据此得到以上三个参数的值。

表 1 计算量比较

参数	计算量	参数	计算量
$T_1=0.15$, $T_2=0.1$	1276	$T_1=0.4$, $T_2=0.3$	2510
$T_1=0.2$, $T_2=0.15$	1242	$T_1=0.4$, $T_2=0.35$	1264
$T_1=0.3$, $T_2=0.2$	562	$T_1=0.5$, $T_2=0.4$	8818
$T_1=0.3$, $T_2=0.25$	556	$T_1=0.6$, $T_2=0.5$	74668

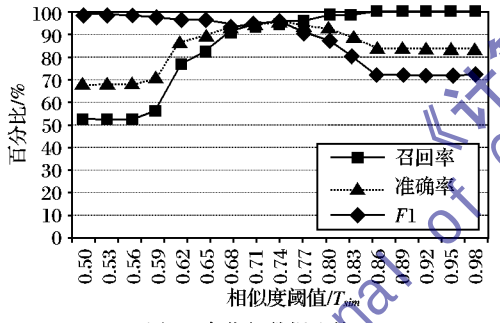

图 3 各指标数据比较

根据已确定的参数,采用逐级聚类算法对测试集进行去重实验,分别统计准确率、召回率和 $F1$ 值,并同 DGHS 算法的相应指标作对比。表 2 为两算法对比结果。

表 2 准确率、召回率和 $F1$ 值比较 %

算法	准确率	召回率	$F1$ 值
逐级聚类算法	95.10	95.72	95.41
DGHS 算法	82.53	84.57	83.54

从表 2 可以看出,逐级聚类算法各指标均优于 DGHS 算法,尤其是准确率提高 12.6%。这是由于 DGHS 算法的 Q-gram相似性度量函数不能有效应用于 Web 整合信息,而基于动态权重的模糊实体匹配策略能够有效识别出企业名称,在属性缺失情况下利用动态权重调整相似度计算,大大提高了匹配准确率,这也是本文算法的最大优点。但由于真实数据噪声较多和编辑距离识别精度有限等原因,逐级聚类算法尚没有达到十分理想的效果,这也是下一步工作的重要研究方向。

2)运行时间对比。

实验对爬取的 60 万条林业企业数据进行了清洗,执行时间的比较如图 4 所示。实验数据显示在大数据量情况下(60 万条),两种算法在不同记录数下的运行时间,随着数据量的增加,逐级聚类算法计算效率获得明显提高。这是由于本文算法在精确去重前的两层聚类,有效将相似记录聚类到相近的位置,且在精确检测相似记录时采用数据并行处理,明显提高了时间效率,缩短了算法的执行时间,而且随着数据量的增加,这种优势会愈加显著。

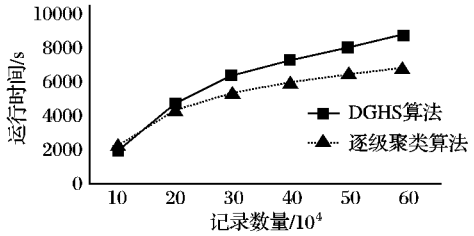

图 4 运行时间比较

5 结语

本文针对数据去重的两个关键问题,结合 Web 信息整合中记录的特点,提出了一种逐级聚类的数据去重方法,通过关键属性分割、Canopy 粗聚类和精确去重等步骤,缩小检索空间;在精确去重阶段提出基于动态权重模糊实体匹配策略,提高了记录匹配精度。最后通过林业 Web 黄页数据去重的实验验证了所提方法在准确率、召回率和时间效率上的明显优势,尤其是准确率提高了 12.6%。今后的工作重点是消除数据噪声,改进记录间相似度的计算方法以提高算法的准确率。

参考文献:

[1] 李广建. 整合研究的几个理论问题[J]. 图书情报工作, 2005, 49(10): 5-10.

[2] 叶焕倬, 吴迪. 相似重复记录清理方法研究综述[J]. 现代图书情报技术, 2010(9): 56-66.

[3] PANSE F, van KEULEN M, de KEIJZER A, *et al*. Duplicate detection in probabilistic data[C]// Proceedings of the 26th International Conference on Data Engineering Workshop. Washington, DC: IEEE Computer Society, 2010: 179-182.

[4] 夏骄雄, 徐俊, 吴耿锋. 数据清理中同体不同源数据的数化算法研究[J]. 计算机工程, 2007, 33(1): 71-73.

[5] HERNANDEZ M A, STOLFO S J. The merge/purge problem for large databases[C]// SIGMOD 1995: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 1995.

[6] 张平, 党选举, 陈皓, 等. 基于熵特征优选分组聚类的相似重复记录检测[J]. 传感器与微系统, 2011, 30(11): 135-137.

[7] 韩京宇, 徐立臻, 董逸生. 一种大数据量的相似记录检测方法[J]. 计算机研究与发展, 2005, 42(12): 2206-2212.

[8] SITAS A, KAPIDAKIS S. Duplicate detection algorithms of bibliographic descriptions[J]. Library Hi-Tech, 2008, 26(2): 287-301.

[9] 唐懿芳, 钟达夫, 严小卫. 基于聚类模式的多数据源记录匹配算法[J]. 小型微型计算机系统, 2005, 26(9): 1546-1550.

[10] 邓攀, 刘功申. 一种高效的倒排索引存储结构[J]. 计算机工程与应用, 2008, 44(31): 149-152.

[11] MCCALLUM A, NIGAM K, UNGAR L H. Efficient clustering of high dimensional data sets with application to reference matching [C]// Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data mining. New York: ACM Press, 2000: 169-178.

[12] 王欢. 林业黄页信息自动分类技术研究[D]. 北京: 北京林业大学, 2012.